

How accurately does the VIVO Harvester reflect actual Clinical and Translational Sciences Award–affiliated faculty member publications?*

Jonathan D. Eldredge, MLS, PhD, AHIP; Philip J. Kroth, MD, MS; Cristina Murray-Krezan, MS; Chad M. Hantak, MS; Edward F. Weagel, MS; Gale G. Hannigan, PhD, AHIP

See end of article for authors' affiliations.

DOI: <http://dx.doi.org/10.3163/1536-5050.103.1.004>

Objective: The research tested the accuracy of the VIVO Harvester software in identifying publications authored by faculty members affiliated with a National Institutes of Health Clinical and Translational Sciences Award (CTSA) site.

Methods: Health sciences librarians created “gold standard” lists of references for the years 2001 to 2011 from PubMed for twenty-five randomly selected investigators from one CTSA site. These gold standard lists were compared to the same twenty-five investigators' reference lists produced by VIVO Harvester. The authors subjected the discrepancies between the lists to sensitivity and specificity analyses.

Results: The VIVO Harvester correctly identified only about 65% of the total eligible PubMed references for the

years 2001–2011 for the CTSA-affiliated investigators. The identified references produced by VIVO Harvester were precise yet incomplete. The sensitivity rate was 0.65, and the specificity rate was 1.00.

Conclusion: While the references produced by VIVO Harvester could be confirmed in PubMed, the VIVO Harvester retrieved only two-thirds of the required references from PubMed. National Institutes of Health CTSA sites will need to supplement VIVO Harvester–produced references with the expert searching skills of health sciences librarians.

Implications: Health sciences librarians with searching skills need to alert their CTSA sites about these deficiencies and offer their skills to advance their sites' missions.

INTRODUCTION

Health sciences librarians primarily serve to connect users with needed health information. The Medical Library Association (MLA) captures this central purpose concisely in its tagline, “Professionals providing quality information for improved health,” and in its *Code of Ethics for Health Sciences Librarianship* [1]. Similarly, health sciences librarians also have connected people with one another for perhaps as long as they existed as a profession. For many years, health sciences librarians have followed John Cotton Dana's advice that librarians be immersed in their user communities [2]. This immersion enables librarians to identify the research interests of their users, which can facilitate connecting researchers with others of similar interests.

The convergence of twenty-first century information technology and librarianship has allowed the augmentation of this researcher-connections role. Librarians now can link researchers with shared subject interests in ways previously difficult to imagine. VIVO software offers a dynamic online directory to facilitate professional networking opportunities for investigators who share similar or complementary research interests [3]. Individual institutions host VIVO software on their servers to enable researchers to network with other researchers within or outside their host institutions. The related VIVO Harvester software automatically populates

investigators' profiles with references ingested from PubMed on to the investigators' institutional hosting of VIVO.

The US National Institutes of Health's (NIH's) Clinical and Translational Sciences Award (CTSA) program currently funds sixty centers at academic health sciences centers [4]. The NIH CTSA collaboration has a keen interest in populating the VIVO database nationwide to facilitate collaboration among investigators. Also, individual institutions want to use the VIVO database to gauge the total research productivity of their faculty members.

VIVO and VIVO Harvester can help researchers identify colleagues who are likely to collaborate in the same or complementary subject areas. This collaboration can also allow them to recruit larger, geographically dispersed study populations. Informatics investigators have discovered that scientists linked together in social networks are far more likely to be “highly successful” compared to solo researchers [5]. Analysis of a twenty-year record in Web of Science, for example, revealed that interdisciplinary, geographically dispersed researchers tend to be the most productive [6].

Health sciences librarians have a stake in the success of resources such as VIVO Harvester. Perhaps for this reason, participants in the MLA Translational Sciences Collaboration Special Interest Group meetings since 2011 have consistently highly ranked their involvement with resources such as VIVO. VIVO also supports the central professional role of librarians in helping researchers make productive interpersonal connections. On a practical level, VIVO Harvester

* Clinical and Translational Sciences Center, University of New Mexico. National Institutes of Health award #UL1TR000041.

Table 1
Sample size stratification schema

Category	Number	Sample number
Basic science	11	3
Clinical	72	16
Pharmacy	10	2
Other	15	4
Totals	108	25

Notes: The category for "Other" consists of 6 faculty members from the College of Nursing, 2 from the library, 3 from psychology, and 4 others from the main campus.

offers a potential alternative to librarians manually producing PubMed searches for each faculty member at their institutions for inclusion in the VIVO database. However, the authors found no published study that evaluates the accuracy of the references produced by the VIVO Harvester.

STUDY OBJECTIVE

This study sought to test the accuracy of VIVO Harvester in identifying correct publications indexed in PubMed authored by faculty members who were affiliated with one CTSA site and were linked to their institution's VIVO instance.

METHODS

During April through July 2012, the authors compiled a complete and accurate list of all 108 faculty members currently employed by the University of New Mexico who had a formal affiliation with its CTSA-funded Clinical and Translational Sciences Center (CTSC). The authors have documented this compilation and selection process elsewhere [7]. Once the authors were highly confident that the affiliated list in this study included all faculty members who belonged in the study population while excluding all others, they stratified this study population by (1) clinical, (2) basic science, (3) pharmacy, or (4) nursing and other faculty members. The four strata were randomly sampled using a web-based randomizer <<http://www.randomizer.org>> to produce a representative and proportionate sample of twenty-five CTSA center-affiliated faculty members for this study. Table 1 displays the study population sampled from the larger population.

Using PubMed for the inclusive years of 2001 to 2011, faculty librarians manually searched for each of the twenty-five CTSC-affiliated faculty members. Once the selected faculty members consented to participate in the study, team members compared these PubMed references to the references on the affiliated faculty members' curricula vitae. The faculty librarians worked with the selected CTSA-affiliated faculty members to reconcile any discrepancies. This process led to the creation of twenty-five "gold standard" lists of references for the years 2001 to 2011 that included all references that appeared in PubMed authored or coauthored by the faculty member.

Table 2
VIVO Harvester adherence to the gold standard

		Gold standard		
		(+)	(-)	
Harvester Results	(+)	419	4	423
	(-)	222	7,417,645	7,417,867
		641	7,417,649	7,418,290
		Sensitivity	0.65	
		Specificity	1.00	

Note: A total of 7,417,290 articles were indexed in PubMed with inclusive publication dates of 2000–2011.

One author ran a "naïve query" of the VIVO database hosted by the institution using the VIVO Harvester software for each of the twenty-five affiliated faculty members. This naïve query represented the kind of search likely to be undertaken by an information technology professional when unaided by health sciences librarians who are well versed in the structure and range of functions of a database such as PubMed. The naïve search consisted of the faculty authors' full last names, coupled with the first letters of their first and middle names and linked to the text-word "New Mexico."

Team members compared the gold standard lists of references that the faculty librarians produced against the VIVO Harvester query results for each of the 25 faculty members. Team members carefully documented the instances when the VIVO Harvester produced either false positives or false negatives, compared to the gold standard PubMed searches. The extent of agreement between the 2 lists of references was determined by calculating the specificity and sensitivity, along with their 95% exact binomial confidence intervals. The number 7,417,290 was used for the denominator since it constituted the total number of items indexed in PubMed during the 2001–2011 study period [8]. The difference of positively identified publications between the gold standard PubMed search results compared to the VIVO Harvester search results was analyzed using the Wilcoxon sign-rank test.

RESULTS

The VIVO Harvester demonstrated a limited ability to accurately identify all pertinent article references from PubMed. The sensitivity of 0.65 (95% exact binomial CI: 0.62, 0.69) shown in Table 2 indicates that VIVO Harvester retrieved only about 65% of the eligible PubMed references. PubMed included 100% (specificity=1.0, 1-sided 97.5% exact binomial lower confidence limit: >0.99) of the references that VIVO Harvester retrieved in its queries. This analysis revealed that the VIVO Harvester displayed high specificity but low sensitivity in identifying authors' publications in PubMed. Stated another way, while most references in the VIVO Harvester query could be found in PubMed, many of the PubMed references in the gold standard PubMed searches did not appear in the VIVO Harvester lists of references (Wilcoxon sign-rank test

for difference $P < 0.0001$). Table 2 displays the sensitivity and specificity analyses.

DISCUSSION

When VIVO Harvester is naïvely queried, it produces largely accurate yet incomplete references ingested from PubMed. This sobering discovery suggests that institutions employing online discovery tools such as VIVO need to recognize such serious limitations. This finding also indicates the long-term need for institutions to elicit the essential guidance of health sciences librarians who are knowledgeable in database searching. Importantly, it signals the short-term need for health sciences librarians to perform manual PubMed searches to supplement VIVO until new resources, such as VIVO Harvester, can dramatically improve their ability to ingest a complete list of publications linked to faculty members.

This study points to the need for developing new resources in the near future. While the study has a modest limitation due to its focus on only one CTSA-funded institution, the results suggest caution when interpreting VIVO Harvester results at all CTSA-funded sites. We intend to develop an institutional inquiry to improve the sensitivity of VIVO Harvester ingestions of PubMed references. In the meantime, institutions using resources such as VIVO Harvester need to understand its limitations.

CONCLUSION

While VIVO Harvester references were found to be accurate, they represented only 65% of the gold standard references. Expert query configurations with the VIVO Harvester software will need to be improved to increase the completeness of its references.

ACKNOWLEDGMENTS

In-kind financial and personnel support was received from the University of New Mexico Health Sciences Library and Informatics Center and University of New Mexico Clinical and Translational Sciences Center (NIH award #UL1TR000041). The authors thank their faculty librarian colleague Richard Carr, AHIP, for providing six of the gold standard searches used in this study.

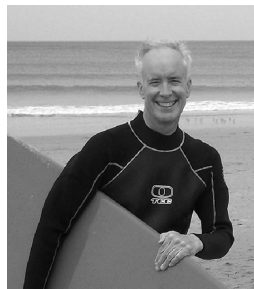
ETHICS ADHERENCE

The University of New Mexico Human Research Protection Office (institutional review board) provided exemption #12-511.

REFERENCES

1. Medical Library Association. Code of ethics for health sciences librarianship [Internet]. Chicago, IL: The Association; 2010 [cited 10 Mar 2013]. <<https://www.mlanet.org/about/ethics.html>>.
2. Dana JC. Hear the other side. In: Hanson CA, ed. Librarian at large: selected writings of John Cotton Dana. Washington, DC: Special Libraries Association; 1991. p. 107–12.
3. VIVO. About [Internet]. [cited 19 Apr 2014]. <<http://www.vivoweb.org/about>>.
4. US National Institutes of Health. Clinical and Translational Sciences Awards (CTSA) [Internet]. The Institutes [cited 16 Apr 2014]. <<https://www.ctsacentral.org>>.
5. Ding Y. Scientific collaboration and endorsement: network analysis of co-authorship and citation networks. *J Informetr*. 2011 Jan 1;5(1):187–203.
6. Contractor N, ed. VIVO Conference 2010 materials “Enabling National Networking of Scientists” 2010 [Internet]. New York, NY; Aug 2010 [cited 19 Apr 2014]. <<http://www.vivoweb.org/conference2010>>.
7. Eldredge JD, Weagel EF, Kroth PJ. Defining and identifying members of a research study population: CTSA-affiliated faculty members. *Hypothesis* [Internet]. 2014 Spring;26(1):5–11 [cited 16 May 2014]. <<http://research.mlanet.org/wp/hypothesis/>>.
8. US National Institutes of Health, National Library of Medicine. Yearly citation totals from 2013 MEDLINE/PubMed baseline [Internet]. The Institutes [cited 10 Mar 2014]. <http://www.nlm.nih.gov/bsd/licensee/2013_stats/2013_Totals.html>.

AUTHORS' AFFILIATIONS



Jonathan D. Eldredge, MLS, PhD, AHIP, jeldredge@unm.edu, Associate Professor, Biomedical Informatics Research, Training and Scholarship, Health Sciences Library and Informatics Center/Family and Community Medicine, University of New Mexico, MSC09 5100, Albuquerque, NM 87131-0001; **Philip J.**

Kroth, MD, MS, pkroth@salud.unm.edu, Director, Biomedical Informatics, Research, Training and Scholarship, and Associate Professor and Section Chief for Clinical Informatics, Divisions of Translational Informatics and General Internal Medicine, Department of Internal Medicine, School of Medicine, University of New Mexico, MSC09 5100, Albuquerque, NM 87131-0001; **Cristina Murray-Krezan, MS**, cmurray-krezan@unm.edu, Research Assistant Professor, Department of Internal Medicine, Division of Epidemiology, Biostatistics, and Preventive Medicine, University of New Mexico, MSC10 0555, Albuquerque, NM 87131-0001; **Chad M. Hantak, MS**, MHantak@unm.edu, Analyst/Programmer, Clinical and Translational Sciences Center, University of New Mexico, MSC 08 4635, Albuquerque, NM 87131-0001; **Edward F. Weagel, MS**, EdWeagel@unm.edu, Biomedical Informatics Manager, Clinical and Translational Sciences Center, University of New Mexico, MSC 08 4635, Albuquerque, NM 87131-0001; **Gale G. Hannigan, PhD, AHIP**, ghannigan@salud.unm.edu, Research Professor, Health Sciences Library and Informatics Center, University of New Mexico, MSC 09 5100, Albuquerque, NM 87131-0001

Received July 2014; accepted August 2014